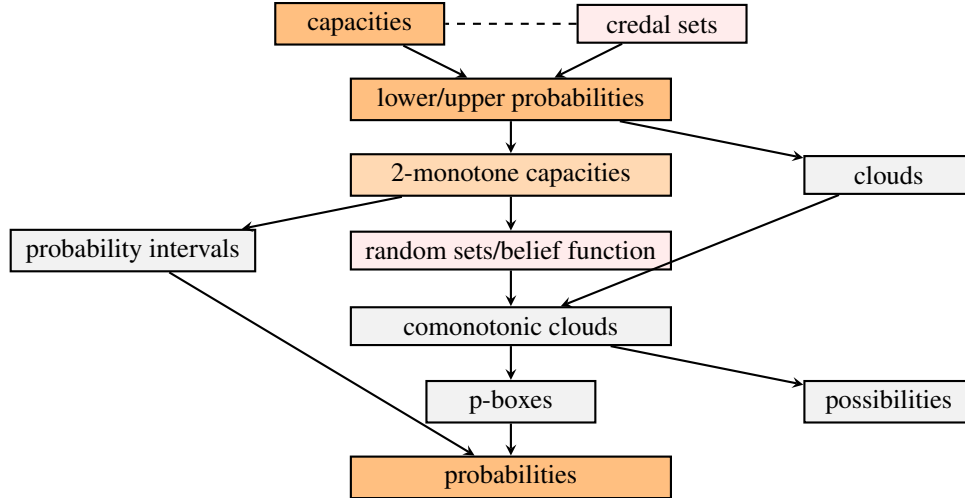


1140	<b>Appendix of <i>Integral Imprecise Probability Metrics</i></b>	
1141	<b>A Some more introduction to Imprecise Probabilistic Machine Learning (IPML)</b>	<b>27</b>
1142	A.1 What is Imprecise Probabilistic Machine Learning actually doing? . . . . .	27
1143	A.2 Credal sets and their use in IPML . . . . .	28
1144	A.3 Belief function/random set and their use in IPML . . . . .	28
1145	<b>B Proofs and derivations</b>	<b>29</b>
1146	B.1 Proof of Lemma 5 . . . . .	29
1147	B.2 Proof of Theorem 6 . . . . .	30
1148	B.3 Proof of Corollary 8 . . . . .	30
1149	B.4 Proof of Proposition 9 . . . . .	30
1150	B.5 Proof of Theorem 10 . . . . .	31
1151	B.6 Proof of Proposition 12 . . . . .	31
1152	B.7 Proof of Remark 14 . . . . .	31
1153	B.8 Proof of Lemma 15 . . . . .	32
1154	B.9 Proof of Theorem 16 . . . . .	32
1155	B.10 Proof of Proposition 18 . . . . .	32
1156	B.11 Proof of Proposition 19 . . . . .	33
1157	B.12 Proof for Theorem 20 . . . . .	33
1158	B.13 Proof of Proposition 21 . . . . .	36
1159	<b>C Further experimental details</b>	<b>37</b>
1160	C.1 Ablation study: Correlation with Generalised Hartley and Entropy Difference. . . . .	37
1161	C.2 Overview of Generalised Hartley measures and Entropy Differences . . . . .	37
1162	C.3 Implementation details . . . . .	39
1163	<b>D IIPM with epsilon-contamination set</b>	<b>39</b>
1164	D.1 Lower Probability Kantorovich Problem . . . . .	40
1165	D.2 Nonparametric Estimator of IIPM with $\epsilon$ -contamination set using kernel distance. . . . .	40



**Figure 2:** A skeleton demonstrating the connection between various uncertainty calculi. “ $A \rightarrow B$ ” means  $A$  generalises  $B$ , meaning that  $B$  is a specific instance of  $A$ . The figure is adopted from Destercke et al. [130] and Hüllermeier and Waegeman [4]. Most of these frameworks generalise classical probability theory. In the main text, we have discussed capacities, lower and upper probabilities, and standard probabilities in detail, with a brief mention of 2-monotone capacities. Additional discussion of credal sets and belief functions is provided in Appendix A. We do not elaborate on the methods shown in grey; for those, we refer readers to existing surveys and review articles.

## A Some more introduction to Imprecise Probabilistic Machine Learning (IPML)

In the main text, we demonstrate the theoretical appeal and practical utility of IIPM primarily through capacities and lower probabilities. While these models are already quite general, we complement this focus by discussing two other mainstream approaches in IPML—credal sets and belief functions—that, although mathematically related, are conceptually motivated in distinct ways. As illustrated in the hierarchy in Figure 2, these models are closely connected to the core ideas of the paper and further support the relevance and broad applicability of the proposed IIPM and MMI framework.

### A.1 What is Imprecise Probabilistic Machine Learning actually doing?

At its core, probabilistic machine learning seeks to construct mathematical models that, through data-driven learning procedures, capture underlying physical or real-world phenomena. For instance, in generative modelling, the objective is to learn the marginal probability distribution that governs the data-generating process. In predictive tasks, instead, the goal is to estimate the conditional distribution of a target variable  $Y$  given an input  $x$ —or its expectation in the case of regression. We often refer to these kinds of natural variation and randomness as aleatoric uncertainty.

Imprecise probabilistic machine learning (IPML) extends this foundation by moving beyond the exclusive use of precise probability models. While classical probability excels at modelling aleatoric uncertainty, IPML incorporates imprecise probability models to account not only for inherent randomness but also for epistemic uncertainty—allowing ambiguity, partial knowledge, and doubt to be explicitly represented within the model. For readers interested in deeper treatments on IP, we recommend Cuzzolin [9] for a comprehensive introduction, Hüllermeier and Waegeman [4, Appendix A], for a concise overview, and Caprio et al. [45, Appendix A], for a discussion on why we should care about imprecision.

In the following, we provide an overview of two other mainstream modelling approaches in IPML: credal set and belief function approaches. We outline the motivations behind these methods and provide some examples of how they are integrated into ML to improve uncertainty quantification and predictive performance.

## 1193 A.2 Credal sets and their use in IPML

1194 Credal sets sit at the top of the hierarchy shown in Figure 2, making them one of the most general  
 1195 constructs in imprecise probability. Many other models can be seen as special cases of credal sets  
 1196 endowed with additional structure. Consequently, there are numerous ways to construct a credal set,  
 1197 depending on the modelling assumptions and information available.

1198 Credal sets are generally understood as some convex set of probability measures  $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$ . Convex-  
 1199 ity can be justified in different ways. In Quasi-Bayesian decision theory [131], it can be shown that  
 1200 rationality axioms proposed for **partial** binary preference naturally leads to a convex set of probability  
 1201 measures, akin to how Savage [132] showed rational binary preference leads to the existence of  
 1202 a unique single probability distribution (paired along with an utility function). Alternatively, in  
 1203 formal epistemology, Williamson [133] put forward the notion of *Chance Calibration*, which is also  
 1204 closely related to Lewis [7]’s *Principal Principle*, which puts into the words of a statistician means,  
 1205 when, according to the current observations, the actual distribution of interest encoding physical  
 1206 phenomena lies within some set  $\{P_1, \dots, P_m\}$  of distributions, but the modeller is indifferent as to  
 1207 which distribution, then rationally, the modeller’s belief, also represented as a distribution, should lie  
 1208 within  $\text{ConvexHull}(\{P_1, \dots, P_m\})$ . This means any distribution in the convex hull is considered a  
 1209 rational belief, thus, the set itself captures the set of rational beliefs, encompassing our imprecision.

1210 In IPML, credal sets are often used to represent either

- 1211 1. Dataset/Distribution-level uncertainty, or
- 1212 2. Predictive uncertainty.

1213 **Case 1.** In the first case, examples include distributionally robust optimisation[134], where a learning  
 1214 algorithm is optimised against the worst-case empirical risk over a credal set—a set of distributions  
 1215 within an  $\epsilon$ -distance from the observed empirical distribution. In the out-of-domain generalisation  
 1216 literature, given observations from multiple source distributions  $\mathbb{P}_1, \dots, \mathbb{P}_m$ , it is commonly assumed  
 1217 that the test-time distribution lies within their convex hull [135, 136], which effectively forms a credal  
 1218 set. Singh et al. [14] made this connection explicit and proposed an algorithm that allows the test-time  
 1219 ambiguity to be resolved without additional training. Caprio et al. [47] developed a learning-theoretic  
 1220 framework for supervised learning under credal sets, while Chau et al. [15] introduced a hypothesis  
 1221 testing procedure for statistically comparing credal sets.

1222 **Case 2.** In the second case, credal sets are used to model epistemic uncertainty in prediction. In  
 1223 credal Bayesian deep learning (CBDL) [45], finitely generated credal sets over priors and likelihoods  
 1224 are combined, by considering all combinatorial applications of Bayes’ rule, to yield a posterior  
 1225 credal set. Wang et al. [80] introduced credal-set interval neural networks, which predict credal sets  
 1226 from probability intervals derived from deterministic interval neural network outputs. Similarly, for  
 1227 classification tasks, Wang et al. [137] proposed defining a predictive credal set as the collection of  
 1228 probability vectors within the simplex that satisfy lower and upper bounds on class probabilities,  
 1229 derived from a set of probabilistic predictors.

## 1230 A.3 Belief function/random set and their use in IPML

1231 Moving down the hierarchy—and skipping lower probabilities and 2-monotone capacities, which are  
 1232 already discussed in the main paper—we briefly describe the roles of random set theory and belief  
 1233 functions.

1234 We focus on finite instance spaces  $\mathcal{X}$ , where random set theory and belief functions coincide. Our  
 1235 exposition follows the terminology of Shafer’s seminal work on The Theory of Evidence [23] and  
 1236 the overview in Cuzzolin [138, Chapter 2.2]. The core philosophy behind belief function theory  
 1237 is its emphasis on representing degrees of support—capturing epistemic uncertainty—rather than  
 1238 specifying how these values are generated, which relates more to aleatoric uncertainty. Perhaps more  
 1239 importantly are the tools developed to combine multiple evidence in a coherent manner.

1240 So, how are belief functions defined? We first introduce a fundamental concept, called basic  
 1241 probability assignments. Let  $\mathcal{X}$  be finite.

1242 **Definition 22.** A basic probability assignment over  $\mathcal{X}$  is a set function  $m : 2^{\mathcal{X}} \rightarrow [0, 1]$  such that

$$m(\emptyset) = 0, \quad \sum_{A \subseteq \mathcal{X}} m(A) = 1.$$

1243 The subsets that have non-zero mass are known as the focal elements within  $2^{\mathcal{X}}$ . Basic probability  
1244 assignment happens in practice when, e.g., sensors have limited precision and can only give results of  
1245 the type “A or B” [139].

1246 Given a mass function  $m$ , which intuitively represents the degree of belief that the true outcome lies  
1247 exactly within the subset  $A \in 2^{\mathcal{X}}$ ,  $m(A)$  quantifies the support assigned to the set  $A$  and no more  
1248 specific subset. From here, we can derive the belief function.

1249 **Definition 23.** The belief function associated with a basic probability assignment  $m : 2^{\mathcal{X}} \rightarrow [0, 1]$  is  
1250 the set function  $\text{Bel} : 2^{\mathcal{X}} \rightarrow [0, 1]$  defined as,

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B).$$

1251 Its conjugate, known as the plausibility function  $\text{Pl}$ , measures the amount of evidence *not against an*  
1252 *event*  $A$  by measuring the following,

$$\text{Pl}(A) = 1 - \text{Bel}(A^c).$$

1253 The theory of evidence is rich and conceptually deep, and a full treatment lies beyond the scope  
1254 of this appendix. However, the context provided should suffice to understand a recent integration  
1255 of belief functions into machine learning. Specifically, Manchingal et al. [140] introduced a new  
1256 class of neural networks, called Random-Set Neural Networks (RS-NNs), for  $K$ -class classification.  
1257 Instead of producing a probability vector with  $K$  outputs—as in standard classifiers—RS-NNs are  
1258 designed to output a basic probability assignment over the  $K$  classes, requiring  $2^K$  output nodes to  
1259 represent belief mass on all subsets of classes. Since this is computationally infeasible for large  $K$ , a  
1260 preprocessing step is introduced to select a subset of focal elements, including the original singleton  
1261 classes and additional relevant subsets. Given the resulting mass function, belief and plausibility  
1262 functions can then be derived for prediction and uncertainty quantification. This principle is further  
1263 extended to convolutional neural networks in Manchingal et al. [78]. As perhaps the first of its kind,  
1264 this random set-based learning paradigm has also been recently adapted to large language models in  
1265 Mubashar et al. [141], offering an explicit mechanism for modelling epistemic uncertainty in LLMs.

## 1266 B Proofs and derivations

1267 This section presents the proofs and derivation in the main text.

### 1268 B.1 Proof of Lemma 5

1269 **Lemma 5.** For lower probability  $\underline{P}$  associated to credal set  $\mathcal{C}$ , we have  $\oint f d\underline{P} \leq \inf_{P \in \mathcal{C}} \int f dP$  for  
1270 any  $f \in C_b(\mathcal{X})$ . When  $\underline{P}$  is 2-monotonic, the inequality becomes an equality.

1271 *Proof.* Let  $\underline{P}$  be the lower probability associated to the credal set  $\mathcal{C}$ . For  $f \in C_b(\mathcal{X})$ , we can write  
1272 the Choquet integral as,

$$\begin{aligned} \oint f d\underline{P} &= \underline{f} + \int_{\underline{f}}^{\bar{f}} \underline{P}(\{f \geq t\}) dt \\ &= \underline{f} + \int_{\underline{f}}^{\bar{f}} \inf_{P \in \mathcal{C}} P(\{f \geq t\}) dt \\ &\leq \underline{f} + \inf_{P \in \mathcal{C}} \int_{\underline{f}}^{\bar{f}} P(\{f \geq t\}) dt \\ &= \inf_{P \in \mathcal{C}} \int f dP. \end{aligned}$$

1273 For 2-monotone  $\underline{P}$ , the results follow from Delbaen [142, Lemma 2]. □

## 1274 B.2 Proof of Theorem 6

1275 **Theorem 6.** *Let  $(\mathcal{X}, d)$  be a metric space. For any capacities  $\nu, \mu \in \mathcal{V}(\mathcal{X})$ , we have  $\oint f d\nu = \oint f d\mu$*   
 1276 *for all  $f \in C_b(\mathcal{X})$ , if and only if  $\nu = \mu$ .*

1277 *Proof.* ( $\implies$ ) Let  $U$  be any open set in  $\mathcal{X}$  and  $F$  the complement. Consider the distance  $d(x, F) =$   
 1278  $\min_{y \in F} d(x, y)$ . For  $n = 1, 2, \dots$ , let  $f_n(x) = \min(1, nd(x, F))$ . Then,  $\sup_{x \in S} |f_n - f| \rightarrow 0$ ,  
 1279 where  $f = \mathbf{1}_U$  is the indicating function. Now, as Choquet integration is continuous with respect to  
 1280 the topology of uniform convergence [19, Proposition C.5(ix)], we have

$$\lim_{n \rightarrow \infty} \oint f_n d\nu = \oint \mathbf{1}_U d\nu = \oint \mathbf{1}_U d\mu.$$

1281 This implies  $\nu(U) = \mu(U)$ . Now, since  $(\mathcal{X}, d)$  is a metric space, for any  $A \in \Sigma_{\mathcal{X}}$ , we can find an  
 1282 increasing sequence of open subsets  $A_1 \subseteq A_2, \dots$  such that  $A_n \uparrow A$ . Since  $\nu, \mu$  are continuous from  
 1283 below, we have  $\lim_{n \rightarrow \infty} \nu(A_n) = \nu(A)$ , but since for any open set  $\nu(A_n) = \mu(A_n)$ , we conclude  
 1284  $\nu(A) = \mu(A)$ , thus  $\nu = \mu$ .

1285 ( $\impliedby$ ) If  $\nu = \mu$ , then it is trivial to see  $\oint f d\nu = \oint f d\mu$  for any  $f \in C_b(S)$ .  $\square$

## 1286 B.3 Proof of Corollary 8

1287 **Corollary 8.** *For any  $P, Q \in \mathcal{P}(\mathcal{X})$  and  $\mathcal{F} \subseteq C_b(\mathcal{X})$ ,  $\text{IIPM}_{\mathcal{F}}(P, Q) = \text{IPM}_{\mathcal{F}}(P, Q)$ .*

1288 *Proof.* For any  $P, Q \in \mathcal{P}(\mathcal{X})$  and  $\mathcal{F} \subseteq C_b(\mathcal{X})$ , we have

$$\begin{aligned} \text{IIPM}_{\mathcal{F}}(P, Q) &= \sup_{f \in \mathcal{F}} \left\{ \left| \oint f dP - \oint f dQ \right| \right\} \\ &= \sup_{f \in \mathcal{F}} \left\{ \left| \int f dP - \int f dQ \right| \right\} \\ &= \text{IPM}_{\mathcal{F}}(P, Q), \end{aligned}$$

1289 since the Choquet integral for additive probability is the Lebesgue integral.  $\square$

## 1290 B.4 Proof of Proposition 9

1291 **Proposition 9.** *For any  $\mathcal{F} \subseteq C_b(\mathcal{X})$ ,  $\text{IIPM}_{\mathcal{F}}$  is a pseudometric on  $\mathcal{V}(\mathcal{X})$ ; it is **non-negative**,*  
 1292 ***symmetric**, and satisfies the **triangle inequality**.*

1293 *Proof.* To prove it is a pseudometric, we need non-negativity, symmetry, and to show triangle  
 1294 inequality.

- 1295 • **Non-negative:** It is obvious that  $\text{IIPM}_{\mathcal{F}}(\nu_1, \nu_2) \geq 0$  for any pair of  $\nu_1, \nu_2 \in \mathcal{V}(\mathcal{X})$
- 1296 • **Symmetric:** Symmetry is also apparent.
- 1297 • **Triangle inequality:** Pick  $\nu_1, \nu_2, \nu_3$  from  $\mathcal{V}(\mathcal{X})$ , then

$$\begin{aligned} \text{IIPM}_{\mathcal{F}}(\nu_1, \nu_2) &= \sup_{f \in \mathcal{F}} \left| \oint f d\nu_1 - \oint f d\nu_2 \right| \\ &= \sup_{f \in \mathcal{F}} \left| \oint f d\nu_1 - \oint f d\nu_3 + \oint f d\nu_3 - \oint f d\nu_2 \right| \\ &\leq \sup_{f \in \mathcal{F}} \left| \oint f d\nu_1 - \oint f d\nu_3 \right| + \sup_{f \in \mathcal{F}} \left| \oint f d\nu_3 - \oint f d\nu_2 \right| \\ &= \text{IIPM}_{\mathcal{F}}(\nu_1, \nu_3) + \text{IIPM}_{\mathcal{F}}(\nu_3, \nu_2). \end{aligned}$$

1298 This concludes the proof. We note that in some literature, pseudometric also requires  $\text{IIPM}_{\mathcal{F}}(\nu, \nu) =$   
 1299 0, and this also trivially holds in our case.  $\square$

## 1300 B.5 Proof of Theorem 10

1301 **Theorem 10.** *Let  $\mathcal{F} \subseteq C_b(\mathcal{X})$  be dense in  $C_b(\mathcal{X})$  with respect to the  $\|\cdot\|_\infty$  norm. Then,  $\text{IIPM}_{\mathcal{F}}$*   
 1302 *metrises the Choquet weak convergence of  $\mathcal{V}(\mathcal{X})$ .*

1303 *Proof.* To show that  $\text{IIPM}_{\mathcal{F}}$  metrises the Choquet weak convergence of  $\mathcal{V}(\mathcal{X})$ , we need to show for  
 1304  $\nu_n, \nu \in \mathcal{V}(\mathcal{X})$ , whenever  $\text{IIPM}_{\mathcal{F}}(\nu_n, \nu) \rightarrow 0$  then  $\nu_n$  converges to  $\nu$  in the Choquet weak sense.

1305 Now pick  $f \in C_b(\mathcal{X})$ , since by assumption  $\mathcal{F}$  is dense in  $C_b(\mathcal{X})$ , there exists  $g \in \mathcal{F}$  satisfying  
 1306  $\|f - g\|_\infty < \epsilon$ . By assumption,  $\text{IIPM}_{\mathcal{F}}(\nu_n, \nu) \rightarrow 0$  means  $|\int f d\nu_n - \int f d\nu| \rightarrow 0$  since  $g \in \mathcal{F}$ .  
 1307 Thus, we have the bound

$$\begin{aligned} \left| \int f d\nu_n - \int f d\nu \right| &\leq \left| \int f d\nu_n - \int g d\nu_n \right| + \left| \int g d\nu_n - \int g d\nu \right| + \left| \int g d\nu - \int f d\nu \right| \\ &\leq 2\epsilon \end{aligned}$$

1308 for all  $f \in C_b(\mathcal{X})$  and  $\epsilon > 0$ , which implies  $\nu_n$  converges to  $\nu$  in the Choquet weak sense.

1309 Proving the other direction is a direct application of the result from Theorem 6 since  $\mathcal{F} \subseteq C_b(\mathcal{X})$ .  $\square$

## 1310 B.6 Proof of Proposition 12

1311 **Proposition 12.** *The Lower Dudley metric metrises Choquet weak convergence on  $\underline{\mathcal{P}}(\mathcal{X})$ .*

1312 *Proof.* The core idea is to show that  $\mathcal{F}_d$  is dense in  $C_b(\mathcal{X})$  and then to apply Theorem 10.

To show denseness, pick any  $f \in C_b(\mathcal{X})$ , consider the sequence

$$f_n(x) = \inf_{y \in \mathcal{X}} \{f(y) + nd(x, y)\}.$$

1313 Pick  $\alpha > 0$  such that  $|f(x)| \leq \alpha$  for all  $x \in \mathcal{X}$ , thus  $|f(x) - f(y)| \leq |f(x)| + |f(y)| \leq \alpha$  for any  
 1314  $x, y \in \mathcal{X}$ . It's clear that  $-\alpha \leq f_n \leq f$  is bounded and  $n$ -Lipschitz continuous. Now we prove  
 1315  $f_n \rightarrow f$  uniformly. Fix  $\epsilon > 0$ , there is  $\delta > 0$  such that  $d(x, y) < \delta$  implies  $|f(x) - f(y)| < \epsilon$  since  
 1316  $f$  is continuous. Then for all  $x \in \mathcal{X}$ , we have

$$\begin{aligned} 0 \leq f(x) - f_n(x) &= f(x) - \inf_{y \in \mathcal{X}} \{f(y) + nd(x, y)\} \\ &= \sup_{y \in \mathcal{X}} \{f(x) - f(y) - nd(x, y)\} \\ &= \sup_{y \in \mathcal{X}, d(x, y) \leq 2\alpha/n} \{f(x) - f(y) - nd(x, y)\}. \end{aligned}$$

1317 If  $n$  is such that  $\frac{2\alpha}{n} < \delta$ , then

$$f(x) - f_n(x) \leq \sup \{\epsilon - nd(x, y) \mid y \in \mathcal{X} \text{ s.t. } d(x, y) \leq 2\alpha/n\} \leq \epsilon$$

1318 which follows that  $\|f - f_n\| \leq \epsilon$ , meaning  $\mathcal{F}_d$  is dense in  $C_b(\mathcal{X})$  in the uniform norm.  $\square$

## 1319 B.7 Proof of Remark 14

1320 **Remark 14.** *Let  $\mathcal{X}$  be finite. For any  $P, Q \in \mathcal{P}(\mathcal{X})$ , we have  $\text{IPM}_{\mathcal{F}_{TV}}(P, Q) = \sup_{A \in \Sigma_{\mathcal{X}}} |P(A) -$   
 1321  $Q(A)| = \sum_{x \in \mathcal{X}} (P(\{x\}) - Q(\{x\}))$ . In contrast, in the imprecise case, there exists  $\underline{P}, \underline{Q} \in \underline{\mathcal{P}}(\mathcal{X})$   
 1322 such that  $\text{IIPM}_{\mathcal{F}_{TV}}(\underline{P}, \underline{Q}) := \sup_{A \in \Sigma_{\mathcal{X}}} |\underline{P}(A) - \underline{Q}(A)| \neq \sum_{x \in \mathcal{X}} (\underline{P}(\{x\}) - \underline{Q}(\{x\}))$ .*

1323 *Proof.* We provide an example based on Montes et al. [143, Example 4] for completeness. Consider  
 1324  $\mathcal{X} = \{x_1, x_2, x_3\}$  and lower probabilities  $\underline{P}_1, \underline{P}_2$  given by

	$\emptyset$	$\{x_1\}$	$\{x_2\}$	$\{x_3\}$	$\{x_1, x_2\}$	$\{x_1, x_3\}$	$\{x_2, x_3\}$	$\mathcal{X}$
1325 $\underline{P}_1$	0	0	0	0	$1/2$	$1/2$	$1/2$	1
$\underline{P}_2$	0	0	0	0	$1/3$	$1/3$	$1/3$	1

1326 Then, we know  $d_1 = \sup_{A \subseteq \mathcal{X}} |\underline{P}_1(A) - \underline{P}_2(A)| = 1/6$ , whereas  $d_2 = \sum_{x \in \mathcal{X}} |\underline{P}_1(\{x\}) -$   
 1327  $\underline{P}_2(\{x\})| = 0$ , therefore  $d_1 \neq d_2$ .  $\square$

## 1328 B.8 Proof of Lemma 15

1329 **Lemma 15.** *Let  $\mathcal{P}_{\epsilon, P}$  be an  $\epsilon$ -contaminated model defined in Equation (2). Then, the associated*  
 1330 *lower probability  $\underline{P}_{\epsilon}$  is given by*

$$\underline{P}_{\epsilon}(A) = \inf_{\tilde{P} \in \mathcal{P}_{\epsilon, P}} \tilde{P}(A) = \begin{cases} (1 - \epsilon)P(A) & \text{for all } A \in \Sigma_{\mathcal{X}} \setminus \{\mathcal{X}\} \\ 1, & \text{for } A = \mathcal{X} \end{cases}$$

1331 *Proof.* This lemma is proved in Walley [8, Section 2.9.2]. □

## 1332 B.9 Proof of Theorem 16

1333 **Theorem 16.** *Let  $\mathcal{F}_W := \{f \in C_b(\mathcal{X}) : \|f\|_L \leq 1\}$  where  $\|f\|_L := \sup_{x, y \in \mathcal{X}} \{|f(x) - f(y)|/c(x, y)\}$ ,*  
 1334 *and  $c$  the transportation cost in a restricted lower probability Kantorovich (RLPK) problem [53,*  
 1335 *Definition 10]. Let  $\underline{P}_{\epsilon}, \underline{Q}_{\epsilon}$  be lower probabilities of the  $\epsilon$ -contaminated models  $\mathcal{P}_{\epsilon, P}$  and  $\mathcal{P}_{\epsilon, Q}$ .*  
 1336 *Then,  $\text{IIPM}_{\mathcal{F}_W}(\underline{P}, \underline{Q})$  coincides with the objective of the RLPK problem, and thus coincides with the*  
 1337 *classical Kantorovich's optimal transport problem involving  $P$  and  $Q$ .*

1338 *Proof.* Our goal is to show that  $\text{IIPM}_{\mathcal{F}_W}(\underline{P}_{\epsilon}, \underline{Q}_{\epsilon})$  coincides with the objective of the restricted lower  
 1339 probability Kantorovich problem. Note that we have,

$$\begin{aligned} \text{IIPM}_{\mathcal{F}_W}(\underline{P}_{\epsilon}, \underline{Q}_{\epsilon}) &= \sup_{f \in \mathcal{F}_W} \left| \int f d\underline{P}_{\epsilon} - \int f d\underline{Q}_{\epsilon} \right| \\ &= \sup_{f \in \mathcal{F}_W} \left| \int_{\underline{f}}^{\bar{f}} [\underline{P}_{\epsilon}(\{f \geq t\}) - \underline{Q}_{\epsilon}(\{f \geq t\})] dt \right| \\ &\stackrel{(\heartsuit)}{=} \sup_{f \in \mathcal{F}_W} \left| \int_{\underline{f}}^{\bar{f}} [\underline{P}_{\epsilon}(\{f > t\}) - \underline{Q}_{\epsilon}(\{f > t\})] dt \right| \\ &= \sup_{f \in \mathcal{F}_W} \left| \int_{\underline{f}}^{\bar{f}} [(1 - \epsilon)P(\{f > t\}) - (1 - \epsilon)Q(\{f > t\})] dt \right| \\ &\stackrel{(\clubsuit)}{\leq} (1 - \epsilon) \sup_{f \in \mathcal{F}_W} \left| \int f dP - \int f dQ \right| \\ &\stackrel{(\spadesuit)}{=} (1 - \epsilon) \inf_{\pi \in \Gamma(P, Q)} \int c(x, y) \pi(dx, dy) \end{aligned}$$

1340 where  $\Gamma(P, Q)$  is the set of joint probability with marginals being  $P$  and  $Q$ . We replaced the inequality  
 1341 with strict inequality in  $\heartsuit$  as in Troffaes and De Cooman [19, Proposition C.3.ii]. In  $\clubsuit$  we used the  
 1342 fact that Choquet integration returns the standard Lebesgue integral when the capacity is a probability  
 1343 measure, and in  $\spadesuit$  we used Kantorovich-Rubinstein theorem [144, Lecture 3], which established the  
 1344 duality between the Kantorovich problem and an IPM formulation using function class  $\mathcal{F}_W$ . This  
 1345 recovered the result from Caprio [53] through the use of our IIPM framework. □

## 1346 B.10 Proof of Proposition 18

1347 **Proposition 18.** *The definition of MMI is equivalent to*

$$\text{MMI}_{\mathcal{F}}(\underline{P}) = \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} 1 - \left( \underline{P}(\{f < t\}) + \underline{P}(\{f \geq t\}) \right) dt \quad (3)$$

1348 *Proof.* To show the result, it follows from the definition that

$$\begin{aligned}\text{MMI}_{\mathcal{F}}(\underline{P}) &= \sup_{f \in \mathcal{F}} \left\{ \int f d\bar{P} - \int f d\underline{P} \right\} \\ &= \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} [\bar{P}(\{f \geq t\}) - \underline{P}(\{f \geq t\})] dt \\ &= \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} [1 - (\underline{P}(\{f < t\}) + \underline{P}(\{f \geq t\}))] dt,\end{aligned}$$

1349 which completes the proof.  $\square$

### 1350 B.11 Proof of Proposition 19

1351 **Proposition 19** (MMI on  $\epsilon$ -contamination set.). *Let  $\underline{P}_\epsilon$  be the lower probability associated with  $\mathcal{P}_{\epsilon, P}$*   
 1352 *and  $\mathcal{F} \subseteq C_b(\mathcal{X})$ . Then  $\text{MMI}_{\mathcal{F}}(\underline{P}_\epsilon) = \epsilon \left( \sup_{f \in \mathcal{F}} \sup_{x, y \in \mathcal{X}} |f(x) - f(y)| \right)$ . For the LTV distance*  
 1353 *with  $\mathcal{F}_{TV} := \{1_A : A \in \Sigma_{\mathcal{X}}\}$ , we have  $\text{MMI}_{\mathcal{F}_{TV}}(\underline{P}_\epsilon) = \sup_{A \in \Sigma_{\mathcal{X}}} \{\bar{P}_\epsilon(A) - \underline{P}_\epsilon(A)\} = \epsilon$ .*

1354 *Proof.* First, it can be shown readily the upper probability model for an  $\epsilon$  contamination set:

$$\bar{P}(A) = \begin{cases} (1 - \epsilon)P(A) + \epsilon & \text{for } A \in \Sigma_{\mathcal{X}} \setminus \emptyset \\ 0 & \text{for } A = \emptyset \end{cases}.$$

1355 Next, we have

$$\begin{aligned}\text{MMI}_{\mathcal{F}}(\underline{P}_\epsilon) &= \sup_{f \in \mathcal{F}} \left\{ \int f d\bar{P} - \int f d\underline{P} \right\} \\ &= \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} [\bar{P}(\{f \geq t\}) - \underline{P}(\{f \geq t\})] dt \\ &= \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} [\epsilon + (1 - \epsilon)P(\{f > t\}) - (1 - \epsilon)P(\{f > t\})] dt \\ &= \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} \epsilon dt \\ &= \epsilon \left( \sup_{f \in \mathcal{F}} [\bar{f} - \underline{f}] \right) = \epsilon \left( \sup_{f \in \mathcal{F}} \sup_{x, y \in \mathcal{X}} |f(x) - f(y)| \right),\end{aligned}$$

1356 which shows the result. For  $\mathcal{F}_{TV}$ , it is straightforward to see that the maximum value the terms in the  
 1357 bracket of the last equation can attain is 1. Therefore,

$$\text{MMI}_{\mathcal{F}_{TV}}(\underline{P}_\epsilon) = \epsilon.$$

1358 This completes the proof.  $\square$

### 1359 B.12 Proof for Theorem 20

1360 To prove our uncertainty measure satisfies the axioms, we need the following useful lemma.

1361 **Lemma 24** (Marginalisation preserves 2-monotonicity.). *If  $\underline{P}$  is a 2-monotone capacity defined on*  
 1362 *the joint measurable space  $(\mathcal{X} \times \mathcal{Y}, \Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{Y}})$ , and let  $\underline{P}_1(\cdot) = \underline{P}(\cdot, \mathcal{Y})$  be the marginal capacity*  
 1363 *on  $(\mathcal{X}, \Sigma_{\mathcal{X}})$  and  $\underline{P}_2(\cdot) = \underline{P}(\mathcal{X}, \cdot)$  be the marginal capacity on  $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ , then  $\underline{P}_1$  and  $\underline{P}_2$  are also*  
 1364 *2-monotone. Marginalisation also preserves 2-alternating.*

1365 *Proof.* Recall the definition of 2-monotonicity, for any  $A, B \in \Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{Y}}$ , we have

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B).$$

1366 Now pick  $A_1, B_1 \in \Sigma_{\mathcal{X}}$ , then consider,

$$\begin{aligned} \underline{P}_1(A_1 \cup B_1) + \underline{P}_1(A_1 \cap B_1) &= \underline{P}((A_1 \cup B_1) \times \mathcal{Y}) + \underline{P}((A_1 \cap B_1) \times \mathcal{Y}) \\ &= \underline{P}((A_1 \times \mathcal{Y}) \cup (B_1 \times \mathcal{Y})) + \underline{P}((A_1 \times \mathcal{Y}) \cap (B_1 \times \mathcal{Y})) \\ &\geq \underline{P}(A_1 \times \mathcal{Y}) + \underline{P}(B_1 \times \mathcal{Y}) \\ &= \underline{P}_1(A_1) + \underline{P}_1(B_1). \end{aligned}$$

1367 This shows that 2-monotonicity holds for  $\underline{P}_1$  and by symmetry, it also holds for  $\underline{P}_2$ . Therefore,  
1368 marginalisation preserves 2-monotonicity. The steps to show marginalisation preserves 2-alternating  
1369 are analogous.  $\square$

1370 **Theorem 20.** For any  $\mathcal{F} \subseteq C_b(\mathcal{X})$ ,  $\text{MMI}_{\mathcal{F}}$  satisfies axioms **A1-A4**. If  $\underline{P}$  is 2-monotonic and  
1371 with  $\mathcal{F} = \mathcal{F}_{12}$  defined above, then  $\text{MMI}_{\mathcal{F}_{12}}$  satisfies axioms **A1-A5**. For **A5**, the subadditivity  
1372 becomes  $\text{MMI}_{\mathcal{F}_{12}} \leq \text{MMI}_{\mathcal{F}_1} + \text{MMI}_{\mathcal{F}_2}$ . If the notion of independence in A6 is taken to be strong  
1373 independence in the sense of Cozman [96], **A6** also holds, and  $\text{MMI}_{\mathcal{F}_{12}} = \text{MMI}_{\mathcal{F}_1} + \text{MMI}_{\mathcal{F}_2}$ .

1374 *Proof.* **Axiom A1.** Starting from Axiom A1 that  $\text{MMI}_{\mathcal{F}}$  is non-negative and bounded. First, recall  
1375 that lower probabilities are super-additive, meaning, for  $A, B \in \Sigma_{\mathcal{X}}$  with  $A \cap B = \emptyset$ , we have

$$\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B).$$

1376 Now let  $A = \{f \geq t\}$  and  $B = \{f < t\}$ , then we have

$$1 \geq \underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\}).$$

1377 This means the integrand in Equation (3) is always non-negative, thus,  $\text{MMI}_{\mathcal{F}}$  is always non-negative.  
1378 To show boundedness, notice that,

$$\begin{aligned} |\text{MMI}_{\mathcal{F}}(\underline{P})| &= \left| \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} [1 - (\underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\}))] dt \right| \\ &\leq \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} |1 - (\underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\}))| dt \\ &\leq \sup_{f \in \mathcal{F}} \int_{\underline{f}}^{\bar{f}} 1 dt \\ &= \sup_{f \in \mathcal{F}} \sup_{x, y \in \mathcal{X}} |f(x) - f(y)| \\ &\leq 2 \sup_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x)|. \end{aligned}$$

1379 As  $f \in C_b(\mathcal{X})$ , by definition, it is a bounded function, thus  $\sup_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x)|$  is bounded.

**Axiom A2.** For continuity, our goal is to show that given a sequence of lower probabilities  $\underline{P}_n$  converges to  $\underline{P}$  in the Choquet weak convergence sense, then  $\text{MMI}_{\mathcal{F}}(\underline{P}_n) \rightarrow \text{MMI}_{\mathcal{F}}(\underline{P})$ . Pick  $\epsilon > 0$ , we know there exists  $n_{\epsilon} \in \mathbb{N}$  such that for all  $n > n_{\epsilon}$ ,

$$\left| \oint f d\underline{P}_n - \oint f d\underline{P} \right| < \epsilon$$

1380 for all  $f \in C_b(\mathcal{X})$ . An immediate result that follows from Troffaes and De Cooman [19, Proposition  
1381 C.5.iv] for upper probabilities is,

$$\left| \oint f d\bar{P}_n - \oint f d\bar{P} \right| = \left| \oint -f d\underline{P}_n - \oint -f d\underline{P} \right| < \epsilon$$

1382 since  $-f \in C_b(\mathcal{X})$ . Now with this  $n$ , we know

$$\begin{aligned}
& |\text{MMI}_{\mathcal{F}}(\underline{P}_n) - \text{MMI}_{\mathcal{F}}(\underline{P})| \\
&= \left| \sup_{f \in \mathcal{F}} \left\{ \int f d\bar{P}_n - \int f d\underline{P}_n \right\} - \sup_{g \in \mathcal{F}} \left\{ \int g d\bar{P} - \int g d\underline{P} \right\} \right| \\
&= \left| \sup_{f \in \mathcal{F}} \left\{ \int f d\bar{P}_n - \int f d\bar{P} + \int f d\bar{P} - \int f d\underline{P} + \int f d\underline{P} - \int f d\underline{P}_n \right\} - \sup_{g \in \mathcal{F}} \left\{ \int g d\bar{P} - \int g d\underline{P} \right\} \right| \\
&\leq \left| \sup_{f \in \mathcal{F}} \left\{ \int f d\bar{P}_n - \int f d\bar{P} \right\} + \sup_{f \in \mathcal{F}} \left\{ \int f d\underline{P} - \int f d\underline{P}_n \right\} + \sup_{g \in \mathcal{F}} \left\{ \int g d\bar{P} - \int g d\underline{P} \right\} - \sup_{g \in \mathcal{F}} \left\{ \int g d\bar{P} - \int g d\underline{P} \right\} \right| \\
&\leq \sup_{f \in \mathcal{F}} \left\{ \left| \int f d\bar{P}_n - \int f d\bar{P} \right| + \left| \int f d\underline{P} - \int f d\underline{P}_n \right| \right\} < 2\epsilon.
\end{aligned}$$

1383 Thus, we have proven continuity of  $\text{MMI}_{\mathcal{F}}$ .

1384 **Axiom A3.** To prove monotonicity, notice if  $\underline{P}$  is setwise dominated by  $\underline{Q}$ , then we have for any  $t$   
1385 and any  $f \in \mathcal{F}$ ,

$$\begin{aligned}
& \underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\}) \leq \underline{Q}(\{f \geq t\}) + \underline{Q}(\{f < t\}) \\
& \implies 1 - (\underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\})) \geq 1 - (\underline{Q}(\{f \geq t\}) + \underline{Q}(\{f < t\})).
\end{aligned}$$

1386 Since the integrand of one integral is always at least as large as the other one, by Troffaes and  
1387 De Cooman [19, Proposition C.5.vi], we have

$$\begin{aligned}
& \int_{\underline{f}}^{\bar{f}} 1 - (\underline{P}(\{f \geq t\}) + \underline{P}(\{f < t\})) dt \geq \int_{\underline{f}}^{\bar{f}} 1 - (\underline{Q}(\{f \geq t\}) + \underline{Q}(\{f < t\})) dt \\
& \implies \text{MMI}_{\mathcal{F}}(\underline{P}) \geq \text{MMI}_{\mathcal{F}}(\underline{Q}).
\end{aligned}$$

1388 **Axiom A4.** Showing probability consistency is almost trivial as any  $P \in \mathcal{P}(\mathcal{X})$  is self-conjugate,  
1389 therefore  $\text{MMI}_{\mathcal{F}}(P) = 0$ .

1390 **Axiom A5.** Recall  $\mathcal{F}_{12}$  is defined as

$$\mathcal{F}_{12} := \{f \in C_b(\mathcal{X}) \mid f(x_1, x_2) = f_1(x_1) + f_2(x_2) \text{ for some } f_1 \in C_b(\mathcal{X}_1), f_2 \in C_b(\mathcal{X}_2)\}$$

1391 also that for axiom A5 we are working with 2-monotonic lower probabilities  $\underline{P}$ , meaning for any  
1392 event  $A, B \in \Sigma_{\mathcal{X}}$ ,

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B). \quad (4)$$

1393 Now with Troffaes and De Cooman [19, Proposition C.7] we know that for 2-monotone capacities,  
1394 the Choquet integral is super-additive, meaning, for  $f \in \mathcal{F}_{12}$

$$\int f d\underline{P} = \int (f_1 + f_2) d\underline{P} \geq \int f_1 d\underline{P} + \int f_2 d\underline{P} = \int f_1 d\underline{P}_1 + \int f_2 d\underline{P}_2.$$

1395 Notice the last equality holds because  $\underline{P}(\{f_1 \geq t\}) = \underline{P}(\{\{x_1 \in \mathcal{X} \mid f_1(x_1) \geq t\} \times \mathcal{X}_2\}) =$   
1396  $\underline{P}_1(\{x_1 \in \mathcal{X}_1 \mid f_1(x_1) \geq t\})$ . Similarly, we can show that for upper probabilities of 2-monotone  
1397 lower probabilities, they are 2-alternating, meaning that

$$\int f d\bar{P} = \int (f_1 + f_2) d\bar{P} \leq \int f_1 d\bar{P} + \int f_2 d\bar{P} = \int f_1 d\bar{P}_1 + \int f_2 d\bar{P}_2$$

1398 Now, combine the two results, we have,

$$\begin{aligned}
& \text{MMI}_{\mathcal{F}_{12}}(\underline{P}) \\
&= \sup_{f \in \mathcal{F}_{12}} \left\{ \int f d\bar{P} - \int f d\underline{P} \right\} \\
&\leq \sup_{f \in \mathcal{F}_{12}} \left\{ \int f_1 d\bar{P}_1 + \int f_2 d\bar{P}_2 - \left( \int f_1 d\underline{P}_1 + \int f_2 d\underline{P}_2 \right) \right\} \\
&\leq \sup_{f \in \mathcal{F}_1} \left\{ \int f d\bar{P}_1 - \int f d\underline{P}_1 \right\} + \sup_{f \in \mathcal{F}_2} \left\{ \int f d\bar{P}_2 - \int f d\underline{P}_2 \right\} \\
&= \text{MMI}_{\mathcal{F}_1}(\underline{P}_1) + \text{MMI}_{\mathcal{F}_2}(\underline{P}_2).
\end{aligned}$$

1399 **Axiom A6.** Now when  $\underline{P}_1$  and  $\underline{P}_2$  are strongly independent, then the credal set associated to  $\underline{P}$ ,  
 1400 denote as  $\mathcal{C}$ , is related to the credal set associated to  $\underline{P}_1$  and  $\underline{P}_2$ , denote as  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , as follows:

$$\mathcal{C}_1 := \{P \in \mathcal{P}(\mathcal{X}) \mid P = P_1 \cdot P_2 \text{ for some } P_1 \in \mathcal{C}_1, P_2 \in \mathcal{C}_2\}.$$

1401 To show that the uncertainty is additive when strong independence holds, we use the following  
 1402 representation of the Choquet integral for 2-monotone lower probabilities,

$$\oint f d\underline{P} = \inf_{P \in \mathcal{C}} \int f dP$$

and similarly for 2-alternating upper probabilities, we have

$$\oint f d\overline{P} = \sup_{P \in \mathcal{C}} \int f dP.$$

1403 Now we can show the result, starting with

$$\begin{aligned} & \text{MMI}_{\mathcal{F}_{12}}(\underline{P}) \\ &= \sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \left\{ \oint f d\overline{P} - \oint f d\underline{P} \right\} \\ &\stackrel{(A)}{=} \sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \left\{ \sup_{P \in \mathcal{C}} \int (f_1 + f_2) dP - \inf_{Q \in \mathcal{C}} \int (f_1 + f_2) dQ \right\} \\ &\stackrel{(B)}{=} \sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \left\{ \sup_{P_1 \in \mathcal{C}_1, P_2 \in \mathcal{C}_2} \left\{ \int f_1 dP_1 + \int f_2 dP_2 \right\} - \inf_{Q_1 \in \mathcal{C}_1, Q_2 \in \mathcal{C}_2} \left\{ \int f_1 dQ_1 - \int f_2 dQ_2 \right\} \right\} \\ &= \sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \left\{ \sup_{P_1 \in \mathcal{C}_1} \int f_1 dP_1 + \sup_{P_2 \in \mathcal{C}_2} \int f_2 dP_2 - \inf_{Q_1 \in \mathcal{C}_1} \int f_1 dQ_1 - \inf_{Q_2 \in \mathcal{C}_2} \int f_2 dQ_2 \right\} \\ &\stackrel{(C)}{=} \sup_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2} \left\{ \oint f_1 d\overline{P}_1 + \oint f_2 d\overline{P}_2 - \oint f_1 d\underline{P}_1 - \oint f_2 d\underline{P}_2 \right\} \\ &= \sup_{f_1 \in \mathcal{F}_1} \left\{ \oint f_1 d\overline{P}_1 - \oint f_1 d\underline{P}_1 \right\} + \sup_{f_2 \in \mathcal{F}_2} \left\{ \oint f_2 d\overline{P}_2 - \oint f_2 d\underline{P}_2 \right\} \\ &= \text{MMI}_{\mathcal{F}_1}(\underline{P}_1) + \text{MMI}_{\mathcal{F}_2}(\underline{P}_2). \end{aligned}$$

1404 In step (A), we used the fact that the Choquet integral of 2-monotone capacities is the lower envelope  
 1405 of the set of expectations with respect to the credal set [142, Lemma 2]. In Step (B), we used the  
 1406 fact that due to strong independence, every element in  $P \in \mathcal{C}$  can be written as a product of some  
 1407  $P_1 \in \mathcal{C}_1$  and  $P_2 \in \mathcal{C}_2$  and the fact that Lebesgue integration is linear, and  $\int f_1(x_1)P(dx_1, dx_2) =$   
 1408  $\int f_1(x_1)P_1(dx_1)$ . Finally, in step (C), we used Lemma 24, which stated that marginals of a 2-  
 1409 monotone capacity remain 2-monotone, therefore, we can write  $\inf_{P_1 \in \mathcal{C}_1} \int f_1 dP_1$  back as a Choquet  
 1410 integral.  $\square$

### 1411 B.13 Proof of Proposition 21

1412 Finally, to prove the result for the upper bound, we first recall an intermediate result from Montes  
 1413 et al. [143, Proposition 8], which provides a construction for the best pessimistic  $\epsilon$ -contamination  
 1414 set approximation to any lower probability  $\underline{P} \in \mathcal{P}(\mathcal{X})$ . Understanding this proposition requires  
 1415 clarifying the concept of dominance and outer approximation.

1416 **Definition 25** (Outer approximation). Let  $\mathcal{X}$  be finite and  $\underline{Q}, \underline{P} \in \mathcal{P}(\mathcal{X})$  two lower probabilities. We  
 1417 say  $\underline{Q}$  is an outer approximation of  $\underline{P}$  if for every event  $A \in 2^{\mathcal{X}}$ ,  $\underline{Q}(A) \leq \underline{P}(A)$ .

The direction of the inequality might be confusing at first, but by realising  $\underline{Q}(A) \leq \underline{P}(A)$  for every  
 event  $A \in 2^{\mathcal{X}}$ , it means the core of  $\underline{Q}$ , i.e.

$$\mathcal{M}(\underline{Q}) = \{Q \in \mathbb{P}(\mathcal{X}) : Q(A) \geq \underline{Q}(A) \quad \forall A \in 2^{\mathcal{X}}\},$$

1418 is at least as large as the core of  $\underline{P}$ , i.e.  $\mathcal{M}(\underline{P}) \subseteq \mathcal{M}(\underline{Q})$ .

1419 **Definition 26** (Undominated outer approximation in  $\mathcal{C}_*(\mathcal{X})$ ). Let  $\underline{P} \in \mathcal{P}(\mathcal{X})$  be a lower probability.  
 1420 Let  $\mathcal{C}_*(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$  be a subspace of which the outer approximation  $\underline{Q}$  resides. We say  $\underline{Q}$  is an  
 1421 undominated outer approximation of  $\underline{P}$  in  $\mathcal{C}_*(\mathcal{X})$  if there exists no other lower probabilities  $\underline{Q}'$  in  
 1422  $\mathcal{C}_*(\mathcal{X})$  such that  $\mathcal{M}(\underline{P}) \subseteq \mathcal{M}(\underline{Q}') \subsetneq \mathcal{M}(\underline{Q})$ .

1423 Now we have all the language to understand the result from Montes et al. [143]. Let  $\mathcal{C}_\epsilon(\mathcal{X})$  denote  
 1424 the space of all  $\epsilon$ -contamination models defined on  $\mathcal{X}$ .

1425 **Proposition 27.** Let  $\underline{P} \in \mathcal{P}(\mathcal{X})$  be a lower probability. Define  $\epsilon \in (0, 1)$  and the probability  $P_0$  by:

$$\epsilon = 1 - \sum_{x \in \mathcal{X}} \underline{P}(\{x\}), \quad P_0(\{x\}) = \frac{\underline{P}(\{x\})}{\sum_{x \in \mathcal{X}} \underline{P}(\{x\})}, \text{ for every } x \in \mathcal{X}.$$

1426 Denote by  $\underline{P}_\epsilon$  the  $\epsilon$ -contaminated model constructed following Lemma 15. Then,  $\underline{P}_\epsilon$  is the unique  
 1427 undominated outer approximation of  $\underline{P}$  in  $\mathcal{C}_\epsilon(\mathcal{X})$ .

1428 With this result, we can now derive the upper bound for  $\text{MMI}_{\mathcal{F}_{TV}}(\underline{P})$ .

1429 **Proposition 21.** Let  $\mathcal{X}$  be finite. For any  $\underline{P} \in \mathcal{P}(\mathcal{X})$ ,  $\text{MMI}_{\mathcal{F}_{TV}}(\underline{P}) \leq 1 - \sum_{x \in \mathcal{X}} \underline{P}(\{x\})$ .

1430 *Proof.* We know for any  $\underline{P}$ ,  $\underline{P}_\epsilon$  constructed as in Proposition 27, is the unique undominated outer  
 1431 approximation of  $\underline{P} \in \mathcal{C}_\epsilon(\mathcal{X})$ . Since it is an outer approximation, meaning that  $\underline{P}_\epsilon(A) \leq \underline{P}(A)$  for  
 1432 all events  $A \in 2^\mathcal{X}$ , therefore by the monotonicity axiom (Axiom 3) in theorem 20, we know for any  
 1433  $\mathcal{F} \subseteq \mathcal{C}_b(\mathcal{X})$ , we have

$$\text{MMI}_{\mathcal{F}}(\underline{P}) \leq \text{MMI}_{\mathcal{F}}(\underline{P}_\epsilon).$$

1434 Now choose  $\mathcal{F} = \mathcal{F}_{TV}$  as in Definition 13, then along with Proposition 19, we have

$$\begin{aligned} \text{MMI}_{\mathcal{F}_{TV}}(\underline{P}) &\leq \text{MMI}_{\mathcal{F}_{TV}}(\underline{P}_\epsilon) \\ &= \epsilon. \\ &= 1 - \sum_{x \in \mathcal{X}} \underline{P}(\{x\}). \end{aligned}$$

1435 This concludes the derivation. □

## 1436 C Further experimental details

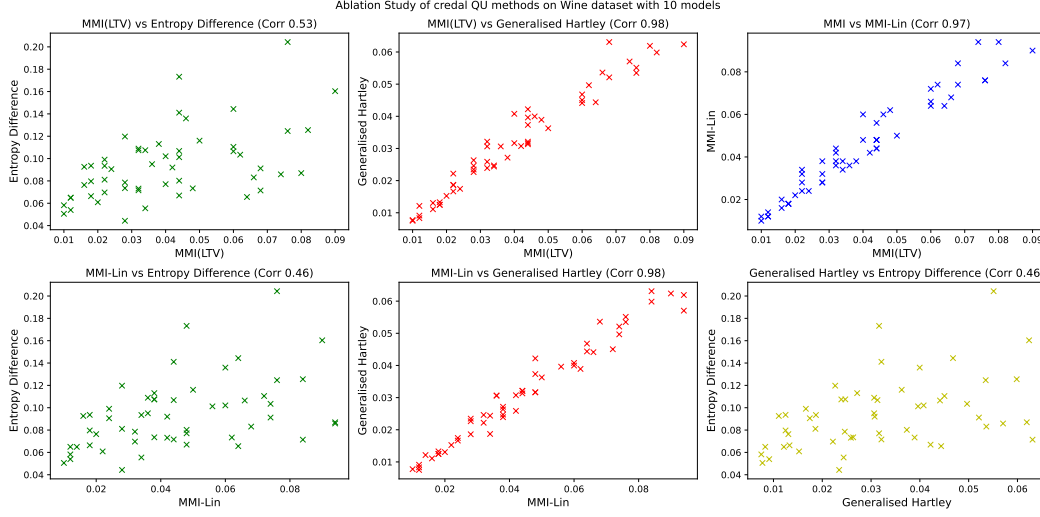
### 1437 C.1 Ablation study: Correlation with Generalised Hartley and Entropy Difference.

1438 In the main text, we observed that the performance of MMI, its linear-time upper bound MMI-Lin,  
 1439 and the generalised Hartley measure are nearly equivalent in the context of selective classification.  
 1440 This raises the question: *are they numerically equivalent?* That is, is generalised Hartley and MMI  
 1441 using the total variation function space, i.e. lower TV, providing the same value? While staring at  
 1442 the equations tells us they are not equivalent, our ablation study suggests that even though they are  
 1443 highly correlated, they are not the same.

1444 To investigate that, we use the UCI wine dataset [145], perform one round of train-test split, train 10  
 1445 random forests models as described in the main text to construct the lower probabilities, and then  
 1446 compute the epistemic uncertainty measurements for the test data. After that, we plot all possible  
 1447 pairwise comparison plots between the methods, i.e. MMI, MMI-Lin, Generalised Hartley, and  
 1448 Entropy difference. We see that MMI and GH are highly correlated but not exactly perfectly correlated.  
 1449 The upper bound is also highly correlated, suggesting empirically (along with experiments in the  
 1450 main text), that this approximation is quite tight. Also, we see that MMI and entropy differences do  
 1451 not correlate that well, which explains the difference in downstream performance in the experiments.

### 1452 C.2 Overview of Generalised Hartley measures and Entropy Differences

1453 We refer the reader to the recent survey by Hoarau et al. [120] on this topic. We hereby provide  
 1454 background on the two methods we compared against in the main text.



**Figure 3:** Comparing the UQ measurements on the withheld test set of the UCI wine dataset [145]. We see that MMI and GH are highly correlated, but not exactly perfectly correlated. The upper bound is also highly correlated, suggesting empirically (along with experiments in the main text), that this approximation is quite tight. Finally, we see that MMI and entropy differences do not correlate that well.

**Generalised Hartley Measure.** Generalised Hartley measure, as the name suggests, is a generalisation of the classical Hartley measure, which is defined as follows:

**Definition 28** (Hartley Measure). *Let  $\mathcal{X}$  be finite. A Hartley measure  $U_H$  is a function  $2^{\mathcal{X}} \mapsto \mathbb{R}$  such that  $U_H(A) = \log_2 |A|$  for  $A \in 2^{\mathcal{X}}$ .*

The Hartley measure can thus be understood as a measure of uncertainty over sets, which can also be viewed as a function on natural numbers. While the expression is simple, Rényi showed that the Hartley measure is the only function mapping natural numbers to the reals that satisfies:

1.  $U_H(mn) = U_H(m) + U_H(n)$  (additivity)
2.  $U_H(m) \geq U_H(m+1)$  (monotonicity)
3.  $U_H(2) = 1$  (normalisation)

These are natural conditions for measuring the amount of uncertainty, or equivalently, information, within a set. The generalised Hartley measure extends this idea of measuring uncertainty in sets in the following way,

**Definition 29** (Generalised Hartley [91]). *Let  $\mathcal{X}$  be finite. Given a lower probability  $\underline{P} \in \mathcal{P}(\mathcal{X})$ , the generalised Hartley  $U_{GH}$  maps  $\underline{P}$  to  $\mathbb{R}$  as follows:*

$$U_{GH}(\underline{P}) = \sum_{A \subseteq \mathcal{X}} m_{\underline{P}}(A) \log_2(|A|),$$

where the mass function  $m_{\underline{P}} : 2^{\mathcal{X}} \rightarrow [0, 1]$  is the Möbius inverse of  $\underline{P}$ , defined as,

$$m_{\underline{P}}(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \underline{P}(B).$$

It is well known that the Möbius inverse is an equivalent representation of a lower probability, in the sense that, once the mass function is known, we can recover  $\underline{P}$  by computing,

$$\underline{P}(A) = \sum_{B \subseteq A} m(B).$$

1473 **Entropy Differences.** To measure aleatoric uncertainty for a given distribution  $P$ , the Shannon  
 1474 entropy is an intuitive solution. The Shannon entropy, or simply entropy, measures the amount of  
 1475 information contained or provided by a source of information.

1476 **Definition 30** (Shannon Entropy). *Let  $\mathcal{X}$  be finite. The Shannon entropy of a distribution  $P$  is*  
 1477 *measured by*

$$U_{Sh}(P) = - \sum_{x \in \mathcal{X}} P(\{x\}) \log_2(P(\{x\})).$$

1478 Now, when given a set of probabilities, i.e. a credal set, to measure the amount of epistemic uncertainty,  
 1479 or in some literature, they call this non-specificity, one natural approach would be to measure the  
 1480 largest difference between the entropies of any two distributions within the set. Formalised below,

1481 **Definition 31** (Entropy Difference). *Let  $\mathcal{X}$  be finite and  $\mathcal{C}$  a credal set. The entropy difference is*  
 1482 *measured by,*

$$\max_{P \in \mathcal{C}} U_{Sh}(P) - \min_{Q \in \mathcal{C}} U_{Sh}(Q).$$

1483 This measure seems intuitive, but violates the monotonicity axioms that a sensible credal UQ measure  
 1484 should satisfy. Consider a credal set  $\mathcal{C}$  and  $P_1, P_2$  the distributions that attained the maximum and  
 1485 minimum of the entropies of the distributions in  $\mathcal{C}$ . Now, enlarge  $\mathcal{C}$  to  $\mathcal{C}'$  by adding distributions that  
 1486 have strictly less entropy than  $P_1$  but more entropy than  $P_2$ , then we have  $\mathcal{C} \subset \mathcal{C}'$ , but the entropy  
 1487 differences will stay the same. This violates the monotonicity axioms that say, when a credal set is  
 1488 strictly larger than the other, then the former should be deemed more epistemically uncertain than the  
 1489 latter.

1490 Nonetheless, the entropy difference is still often used in practice as it is simple to compute, and it  
 1491 can be used to decompose 'total uncertainty' into 'aleatoric' and 'epistemic' components. See for  
 1492 example in Abellán et al. [97], the author defined,

$$\underbrace{\max_{P \in \mathcal{C}} U_{Sh}(P)}_{\text{Total Uncertainty}} = \underbrace{\min_{Q \in \mathcal{C}} U_{Sh}(Q)}_{\text{Aleatoric Uncertainty}} + \underbrace{\max_{P \in \mathcal{C}} U_{Sh}(P) - \min_{Q \in \mathcal{C}} U_{Sh}(Q)}_{\text{Epistemic Uncertainty}}.$$

### 1493 C.3 Implementation details

1494 We provide full experimental details here, which were abbreviated in Section 6 due to space constraints.  
 1495 The experiments were executed on a machine with 8 vCPUs, 30 GB memory, with a NVIDIA V100  
 1496 GPU.

1497 We evaluate the performance of Maximum Mean Imprecision using a selective classification task,  
 1498 following the setup in Shaker and Hüllermeier [54] and Shaker and Hüllermeier [77]. Each dataset is  
 1499 split into training and test sets. For tabular datasets (the Obesity dataset from UCI and Digits dataset  
 1500 from Sci-kit learn), we train 10 random forests with randomly chosen hyperparameters (e.g., tree  
 1501 depth) on the same training set, and evaluate them on the test set. For image datasets (CIFAR10 and  
 1502 CIFAR100), we use 10 pretrained neural networks per task, available at [https://github.com/](https://github.com/chenyaofu/pytorch-cifar-models)  
 1503 [chenyaofu/pytorch-cifar-models](https://github.com/chenyaofu/pytorch-cifar-models). These models were trained using PyTorch's default CIFAR  
 1504 training sets, and we evaluate the standard CIFAR test sets by dividing them into 10 buckets to  
 1505 introduce variability.

1506 For both tabular and image data, we use the centroid of the credal set as the predictor, similar to  
 1507 standard ensemble methods. The corresponding lower probability is computed by evaluating the most  
 1508 pessimistic likelihood across the set of predictions for each possible outcome.

## 1509 D IIPM with epsilon-contamination set

1510 In this section of the appendix, we continue from Section 3 and dive deeper into the connections  
 1511 between IIPM and the epsilon contamination model—a popular class of imprecise models.

## 1512 D.1 Lower Probability Kantorovich Problem

1513 In this subsection, we provide the background on a recently considered problem called the lower  
 1514 probability Kantorovich problem, proposed in Caprio [53]. We start by reviewing what a classical  
 1515 Kantorovich problem is.

1516 **Classical Kantorovich problem.** Let  $P, Q \in \mathcal{P}(\mathcal{X})$  be two probability measures on  $\mathcal{X}$ . Let  
 1517  $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  be a measurable cost function that gives us the cost of moving on unit of probability  
 1518 mass from the first argument to the other. Then the classical Kantorovich problem is the following  
 1519 optimisation problem,

$$\arg \inf_{\alpha \in \Gamma(P, Q)} \left\{ \int c(x, z) d\alpha(x, z) \right\}$$

1520 where  $\Gamma(P, Q)$  is the set of all joint probability measures whose marginals are  $P$  and  $Q$ .  $\alpha$  is also  
 1521 denote as the *transportation plan*, and this is also famously known as the *optimal transport* problem.

1522 **Lower Probability Kantorovich problem.** Caprio [53] consider the following research question,

1523 *What does Kantorovich's problem look like, when instead of transporting probability measures, we*  
 1524 *transport lower probabilities?*

1525 As his answer, Caprio [53] provided the following characterisation of the problem,

1526 **Definition 32** (Lower Probability Kantorovich's OT problem; LPK). *Let  $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  be a*  
 1527 *Borel measurable cost function. Given lower probabilities  $\underline{P}$  and  $\underline{Q}$  on  $\mathcal{X}$ , we want to find the joint*  
 1528 *lower probability alpha on  $\mathcal{X} \times \mathcal{X}$  that solves the following optimisation problem*

$$\arg \inf_{\underline{\alpha} \in \Gamma(\underline{P}, \underline{Q})} \left\{ \int_{\mathcal{X} \times \mathcal{X}} c(x, z) d\underline{\alpha}(x, z) \right\},$$

1529 where  $\Gamma(\underline{P}, \underline{Q})$  is the collection of all joint lower probabilities on  $\mathcal{X} \times \mathcal{X}$  whose marginals on  $\mathcal{X}$  are  
 1530  $\underline{P}$  and  $\underline{Q}$ , respectively.

1531 While this might seem like a straightforward extension from the classical formulation, it is important  
 1532 to note that for imprecise probability theory, there is no unique way to perform conditioning, which  
 1533 means extra care has to be taken into defining  $\Gamma(\underline{P}, \underline{Q})$ . In particular, they focus on a subset of the  
 1534 joint lower probabilities constructed from using geometric conditioning, and called the corresponding  
 1535 LPK problem restricted to such conditioning set the restricted LPK problem.

1536 Later on in Theorem 11 of [53], they managed to show that the restricted LPK problem coincides  
 1537 exactly with the classical Kantorovich for *epsilon*-contaminated sets. We managed to recover this  
 1538 result in our Theorem 16 without needing to consider any specific type of conditioning. In the future,  
 1539 we will investigate how could our result complements to their theory, perhaps allow them to consider  
 1540 other types of conditioning operations in IP.

## 1541 D.2 Nonparametric Estimator of IIPM with $\epsilon$ -contamination set using kernel distance.

1542 Now consider  $\epsilon, \delta \in (0, 1)$  two contamination levels, and distributions  $P, Q \in \mathcal{P}(\mathcal{X})$  which we have  
 1543 i.i.d samples from. Specifically, let  $X_1, \dots, X_n \stackrel{iid}{\sim} P$  and  $Z_1, \dots, Z_m \stackrel{iid}{\sim} Q$  be random variables  
 1544 taking values in  $\mathcal{X}$ . We are interested in quantifying the difference between the  $\epsilon$ -contaminated model  
 1545 of  $P$ , i.e.  $\underline{P}_\epsilon$ , with respect to the  $\delta$ -contaminated model  $Q$ , i.e.  $\underline{Q}_\delta$ .

1546 **A short overview of kernel distances (MMD).** For generic spaces  $\mathcal{X}$ , with iid samples from  
 1547  $P$  and  $Q$ , a popular class of non-parametric discrepancy estimator is the maximum mean discrep-  
 1548 ancancy (MMD) [64, 68]. Specifically, pick a kernel function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  and consider the uniform  
 1549 ball in the corresponding reproducing kernel Hilbert space (RKHS), i.e.  $\mathcal{F}_k = \{f \in \mathcal{H}_k \text{ s.t. } \|f\|_k =$   
 1550  $1\}$ , where  $\|\cdot\|_k$  stands for the RKHS norm. The MMD can be expressed as an IPM with respect to

1551 function class  $\mathcal{F}_k$ ,

$$\begin{aligned}
\text{MMD}_k(P, Q) &= \text{IPM}_{\mathcal{F}_k}(P, Q) \\
&= \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| \int f dP - \int f dQ \right| \right\} \\
&\stackrel{(A)}{=} \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| \langle f, \int k(X, \cdot) dP(X) \rangle - \langle f, \int k(X, \cdot) dQ(X) \rangle \right| \right\} \\
&= \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| \left\langle f, \int k(X, \cdot) dP(X) - \int k(X, \cdot) dQ(X) \right\rangle \right| \right\} \\
&\stackrel{(B)}{=} \left\| \int k(X, \cdot) dP(X) - \int k(X, \cdot) dQ(X) \right\|_k \\
&\stackrel{(C)}{=} \|\mu_P - \mu_Q\|_k
\end{aligned}$$

where in step (A) we first use the reproducing property of RKHS functions,  $(f(x) = \langle f, k(x, \cdot) \rangle)$ , and then use the linearity of the inner product to ‘push’ the expectation inside. In step (B) we use the fact that inner product is maximised when the two vectors align, so we pick the unit-norm function to be

$$f = \frac{\int f(k(X, \cdot) dP(X) - \int k(X, \cdot) dQ(X))}{\left\| \int f(k(X, \cdot) dP(X) - \int k(X, \cdot) dQ(X) \right\|}.$$

1552 Finally, in step (C), we simply write the (Bochner) integral of the feature representation into a more  
 1553 familiar-looking expression, called the kernel mean embedding [146, 147]. This simple expression  
 1554 facilitates further simplification, i.e.

$$\begin{aligned}
\text{MMD}_k(P, Q)^2 &= \|\mu_P - \mu_Q\|_k^2 \\
&= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle \\
&\stackrel{(D)}{=} \mathbb{E}_{X, X' \sim P} [k(X, X')] - 2\mathbb{E}_{X \sim P, Z \sim Q} [k(X, Z)] + \mathbb{E}_{Z, Z' \sim Q} [k(Z, Z')].
\end{aligned}$$

1555 Meaning that we don’t need a parametric assumption on how the data is distributed, as long as we  
 1556 have a way to define similarity between instances through a kernel function, we can estimate the  
 1557 difference between the distributions based on their samples by estimating the expectations in step  
 1558 D. Furthermore, for a wide class of kernels, known as charactersitic kernels, the mapping from  
 1559  $P \mapsto \int k(X, \cdot) dP(X)$  is injective, thus  $\text{MMD}_k$  is a proper metric between probability distributions.  
 1560 Owing to its simplicity, kernel mean embeddings and MMDs have been used to tackle a broad range  
 1561 of statistical tasks, ranging from hypothesis testing [68] to parameter estimation [148, 149], causal  
 1562 inference [150, 151], feature attribution [152, 43], and learning on distributions [153, 154].

1563 **Generalising to contamination sets.** In general, given a lower probability  $\underline{P}$ , the notion of samples  
 1564 from  $\underline{P}$  is ill-defined as lower probability often is used to encode subjective assessment rather than  
 1565 describing the data-generating process. As such, devising a sample-based estimator for the Choquet  
 1566 integral, akin to Monte Carlo estimation for the Lebesgue integral, is not yet possible. Nonetheless,  
 1567 in the case of utilising lower probability constructed through an epsilon contamination model, this is  
 1568 possible.

1569 Recall  $\epsilon, \delta \in (0, 1)$  are two contamination level, with  $\underline{P}_\epsilon$  and  $\underline{Q}_\delta$  the corresponding contaminated  
 1570 models. We are now interested in quantifying the difference between this two lower probabilities using  
 1571 the IIPM framework through the kernel distances. As before, pick  $\mathcal{F}_k = \{f \in \mathcal{H}_k \text{ s.t. } \|f\|_k = 1\}$ ,  
 1572 then we have

$$\begin{aligned}
\text{IIPM}_{\mathcal{F}_k}(\underline{P}_\epsilon, \underline{Q}_\delta) &= \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| \oint f d\underline{P}_\epsilon - \oint f d\underline{Q}_\delta \right| \right\} \\
&\stackrel{(i)}{=} \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| \int_{\underline{f}}^{\overline{f}} \left( \underline{P}_\epsilon(\{f > t\}) - \underline{Q}_\delta(\{f > t\}) \right) dt \right| \right\} \\
&\stackrel{(ii)}{=} \sup_{f \in \mathcal{H}_k; \|f\|_k=1} \left\{ \left| (1 - \epsilon) \int f dP - (1 - \delta) \int f dQ \right| \right\} \\
&\stackrel{(iii)}{=} \|(1 - \epsilon)\mu_P - (1 - \delta)\mu_Q\|_k
\end{aligned}$$

1573 where in step (i) we simply expand our the definition of Choquet integral. In step (ii), we follow  
 1574 Lemma 15 and the proof of Theorem 16 to express the Choquet integral now as a weighted Lebesgue  
 1575 integral. In step (iii), we follow the derivations of standard MMD provided in the previous paragraph.  
 1576 Subsequently, the square of  $\text{IIPM}_{\mathcal{F}_k}(\underline{P}_\epsilon, \underline{Q}_\delta)$  can be expressed as,

$$\begin{aligned} \text{IIPM}_{\mathcal{F}_k}(\underline{P}_\epsilon, \underline{Q}_\delta) &= (1 - \epsilon)^2 \mathbb{E}_{X, X' \sim P}[k(X, X')] \\ &\quad - 2(1 - \epsilon)(1 - \delta) \mathbb{E}_{X \sim P, Z \sim Q}[k(X, Z)] + (1 - \delta)^2 \mathbb{E}_{Z, Z' \sim Q}[k(Z, Z')]. \end{aligned}$$

1577 This allows us to then construct a non-parametric (unbiased) estimator of (the square of  
 1578  $\text{IIPM}_{\mathcal{F}_k}(\underline{P}_\epsilon, \underline{Q}_\delta)$ ) as follows,

$$\begin{aligned} \widehat{\text{IIPM}_{\mathcal{F}_k}^2}(\underline{P}_\epsilon, \underline{Q}_\delta) &= (1 - \epsilon)^2 \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n k(X_i, X_j) \\ &\quad - 2(1 - \epsilon)(1 - \delta) \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Z_j) + (1 - \delta)^2 \sum_{i=1}^m \sum_{j=1, j \neq i}^m k(Z_i, Z_j) \end{aligned}$$

1579 Due to project scope, we did not further investigate the concrete applications of such a non-parametric  
 1580 worst-case probability discrepancy estimator, but in future work, we will explore its use case in robust  
 1581 two-sample testing, akin to Schrab and Kim [155], or in generative model training, akin to Li et al.  
 1582 [128].

## References

- [1] Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.
- [2] Donald Rumsfeld. *Known and unknown: a memoir*. Penguin, 2011.
- [3] Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.
- [4] Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. arXiv:1910.09457 [cs, stat].
- [5] Frank Hyneman Knight. *Risk, uncertainty and profit*, volume 31. Houghton Mifflin, 1921.
- [6] John Maynard Keynes. *A treatise on probability*. 1921.
- [7] David Lewis. A subjectivist’s guide to objective chance. In *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 267–297. Springer, 1980.
- [8] Peter Walley. *Statistical reasoning with imprecise probabilities*, volume 42. Springer, 1991.
- [9] Fabio Cuzzolin. *The geometry of uncertainty: the geometry of imprecise probabilities*. Springer Nature, 2020.
- [10] Alan Hájek. Interpretations of Probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL <https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/>.
- [11] Joseph Bertrand. *Calcul des probabilités*. Gauthier-Villars, 1889.
- [12] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [13] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation, January 2013. URL <http://arxiv.org/abs/1301.2115>. arXiv:1301.2115 [cs, stat].
- [14] Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. *arXiv preprint arXiv:2404.04669*, 2024.
- [15] Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2025.
- [16] Shireen Kudukkil Manchingal and Fabio Cuzzolin. Position: Epistemic artificial intelligence is essential for machine learning models to know when they do not know. *arXiv preprint arXiv:2505.04950*, 2025.
- [17] Vincent Fortuin, Mohammad Emtiyaz Khan, Mark van der Wilk, Zoubin Ghahramani, and Katharine Fisher. Rethinking the role of bayesianism in the age of modern ai (dagstuhl seminar 24461). *Dagstuhl Reports*, 14(11):40–59, 2025.
- [18] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.
- [19] Matthias CM Troffaes and Gert De Cooman. *Lower previsions*. John Wiley & Sons, 2014.

- 1627 [20] Thomas Augustin, Frank P. A. Coolen, Gert De Cooman, and Matthias C. M. Troffaes, editors.  
1628 *Introduction to imprecise probabilities*. Wiley series in probability and statistics. Wiley,  
1629 Hoboken, NJ, 2014. ISBN 978-0-470-97381-3.
- 1630 [21] JFC Kingman. G. matheron, random sets and integral geometry. 1975.
- 1631 [22] Michio Sugeno. Theory of fuzzy integrals and its applications. *Doctoral Thesis, Tokyo Institute*  
1632 *of Technology*, 1974.
- 1633 [23] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press,  
1634 1976.
- 1635 [24] Jonathan Baron. Second-order probabilities and belief functions. *Theory and Decision*, 23(1):  
1636 25–36, 1987.
- 1637 [25] Haim Gaifman. A theory of higher order probabilities. In *Theoretical aspects of reasoning*  
1638 *about knowledge*, pages 275–292. Elsevier, 1986.
- 1639 [26] Didier Dubois and Henri Prade. *Théorie des possibilités*, 1985.
- 1640 [27] Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and*  
1641 *chance*. MIT press, 1980.
- 1642 [28] Michele Caprio and Teddy Seidenfeld. Constriction for sets of probabilities. In Enrique  
1643 Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi, editors, *Proceedings of*  
1644 *the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*,  
1645 volume 215 of *Proceedings of Machine Learning Research*, pages 84–95. PMLR, 11–14 Jul  
1646 2023. URL <https://proceedings.mlr.press/v215/caprio23b.html>.
- 1647 [29] Michele Caprio, Yusuf Sale, Eyke Hüllermeier, and Insup Lee. A Novel Bayes’ Theorem for  
1648 Upper Probabilities. In Fabio Cuzzolin and Maryam Sultana, editors, *Epistemic Uncertainty*  
1649 *in Artificial Intelligence*, pages 1–12, Cham, 2024. Springer Nature Switzerland.
- 1650 [30] G. Choquet. *Théorie des capacités*. *Ann. Inst. Fourier* 5 (1953/1954) 131–292., 1953.
- 1651 [31] Thierry Denoeux. A neural network classifier based on dempster-shafer theory. *IEEE Trans-*  
1652 *actions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150,  
1653 2000.
- 1654 [32] Marco Zaffalon. Exact credal treatment of missing data. *Journal of statistical planning and*  
1655 *inference*, 105(1):105–122, 2002.
- 1656 [33] Michele Caprio, David Stutz, Shuo Li, and Arnaud Doucet. Conformalized credal regions for  
1657 classification with ambiguous ground truth. *arXiv preprint arXiv:2411.04852*, 2024.
- 1658 [34] Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet.  
1659 Credal two-sample tests of epistemic ignorance. *arXiv preprint arXiv:2410.12921*, 2024.
- 1660 [35] Mira Jürgens, Thomas Mortier, Eyke Hüllermeier, Viktor Bengs, and Willem Waegeman. A  
1661 calibration test for evaluating set-based epistemic uncertainty representations. *arXiv preprint*  
1662 *arXiv:2502.16299*, 2025.
- 1663 [36] Christian Fröhlich and Robert C Williamson. Scoring rules and calibration for imprecise  
1664 probabilities. *arXiv preprint arXiv:2410.23001*, 2024.
- 1665 [37] Anurag Singh, Siu Lun Chau, and Krikamol Muandet. Truthful elicitation of imprecise  
1666 forecasts. *arXiv preprint arXiv:2503.16395*, 2025.
- 1667 [38] David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil,  
1668 and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *arXiv preprint*  
1669 *arXiv:2307.09302*, 2023.
- 1670 [39] Michele Caprio, Yusuf Sale, and Eyke Hüllermeier. Conformal prediction regions are imprecise  
1671 highest density regions. *arXiv preprint arXiv:2502.06331*, 2025.

- [40] Fabio Cuzzolin and Ruggero Frezza. Evidential reasoning framework for object tracking. In *Telemanipulator and Telepresence Technologies VI*, volume 3840, pages 13–24. SPIE, 1999.
- [41] Eleonora Giunchiglia, Mihaela Cătălina Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. Road-r: the autonomous driving dataset with logical requirements. *Machine Learning*, 112(9):3261–3291, 2023.
- [42] Jack Liell-Cock and Sam Staton. Compositional imprecise probability: A solution from graded monads and markov categories. 9(POPL), 2025. doi: 10.1145/3704890. URL <https://doi.org/10.1145/3704890>.
- [43] Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36:50769–50795, 2023.
- [44] Lev Utkin, Andrei Konstantinov, Kirill Vishniakov, and Igor Ilin. Imprecise shap as a tool for explaining the class probability distributions under limited training data. In *Digital Systems and Information Technologies in the Energy Sector*, pages 369–389. Springer, 2025.
- [45] Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal bayesian deep learning. *arXiv e-prints*, pages arXiv–2302, 2023.
- [46] Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, Hans Hallez, et al. Credal deep ensembles for uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:79540–79572, 2024.
- [47] Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal Learning Theory, February 2024. arXiv:2402.00957 [cs, stat].
- [48] Fabio G Cozman. Credal networks. *Artificial intelligence*, 120(2):199–233, 2000.
- [49] Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, and David Huber. Approximating counterfactual bounds while fusing observational, biased and randomised data sources. *International Journal of Approximate Reasoning*, 162:109023, 2023.
- [50] Souradeep Dutta, Michele Caprio, Vivian Lin, Matthew Cleaveland, Kuk Jin Jang, Ivan Ruchkin, Oleg Sokolsky, and Insup Lee. Distributionally Robust Statistical Verification with Imprecise Neural Networks. *PMLR (Accepted to HSCC 2025)*, 2024.
- [51] Pengyuan Lu, Michele Caprio, Eric Eaton, and Insup Lee. IBCL: Zero-shot Model Generation for Task Trade-offs in Continual Learning. *arXiv preprint arXiv:2305.14782*, 2024.
- [52] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.
- [53] Michele Caprio. Optimal transport for  $\epsilon$ -contaminated credal sets. *PMLR (accepted to ISIPTA 2025)*, 2025. URL <https://arxiv.org/abs/2410.03267>.
- [54] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and Epistemic Uncertainty with Random Forests, January 2020. URL <http://arxiv.org/abs/2001.00893>. arXiv:2001.00893 [cs, stat].
- [55] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.
- [56] Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.
- [57] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.
- [58] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

- 1719 [59] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley*  
1720 *symposium on mathematical statistics and probability, volume 1: contributions to the theory*  
1721 *of statistics*, volume 4, pages 547–562. University of California Press, 1961.
- 1722 [60] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG  
1723 Lanckriet. On the empirical estimation of integral probability metrics. 2012.
- 1724 [61] Vladimir Mikhailovich Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*,  
1725 28(2):264–287, 1983.
- 1726 [62] Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian  
1727 processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.
- 1728 [63] Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Man-*  
1729 *agement science*, 6(4):366–422, 1960.
- 1730 [64] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A  
1731 kernel method for the two-sample-problem. *Advances in Neural Information Processing*  
1732 *Systems*, 19:513–520, 2006.
- 1733 [65] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of  
1734 dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical*  
1735 *statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University  
1736 of California Press, 1972.
- 1737 [66] Aad W Van Der Vaart, Jon A Wellner, Aad W van der Vaart, and Jon A Wellner. *Weak*  
1738 *convergence*. Springer, 1996.
- 1739 [67] Richard M Dudley. *Real analysis and probability*. 2002.
- 1740 [68] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander  
1741 Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773,  
1742 2012.
- 1743 [69] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy  
1744 gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.
- 1745 [70] Henry E Kyburg Jr. Bayesian and non-bayesian evidential updating. *Artificial intelligence*, 31  
1746 (3):271–293, 1987.
- 1747 [71] Vladimir Vovk and Glenn Shafer. Game-theoretic probability. *Introduction to Imprecise*  
1748 *Probabilities*, pages 114–134, 2014.
- 1749 [72] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.
- 1750 [73] Glenn Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1:330–331,  
1751 1992.
- 1752 [74] Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. Ergodic theorems for  
1753 lower probabilities. *Proceedings of the American Mathematical Society*, 144(8):3381–3396,  
1754 2016.
- 1755 [75] Francesca Mangili. A prior near-ignorance gaussian process model for nonparametric regres-  
1756 sion. *International Journal of Approximate Reasoning*, 78:153–171, 2016.
- 1757 [76] Giorgio Corani and Andrea Mignatti. Credal model averaging for classification: representing  
1758 prior ignorance and expert opinions. *International Journal of Approximate Reasoning*, 56:  
1759 264–277, 2015.
- 1760 [77] Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification:  
1761 Bayesian versus credal inference. In *PROCEEDINGS 31. WORKSHOP COMPUTATIONAL*  
1762 *INTELLIGENCE*, volume 25, page 63, 2021.

- 1763 [78] Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar,  
1764 and Fabio Cuzzolin. Random-set convolutional neural network (rs-cnn) for epistemic deep  
1765 learning. 2023.
- 1766 [79] Jonathan Sadeghi, Marco De Angelis, and Edoardo Patelli. Efficient training of interval neural  
1767 networks for imprecise training data. *Neural Networks*, 118:338–351, 2019.
- 1768 [80] Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David  
1769 Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estima-  
1770 tion in classification tasks. *Neural Networks*, page 107198, 2025.
- 1771 [81] Yasuo Narukawa. Inner and outer representation by choquet integral. *Fuzzy Sets and Systems*,  
1772 158(9):963–972, 2007.
- 1773 [82] Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing  
1774 weak convergence with maximum mean discrepancies. *Journal of Machine Learning  
1775 Research*, 24(184):1–20, 2023.
- 1776 [83] Ding Feng and Hung T Nguyen. Choquet weak convergence of capacity functionals of random  
1777 sets. *Information Sciences*, 177(16):3239–3250, 2007.
- 1778 [84] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic  
1779 kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7),  
1780 2011.
- 1781 [85] Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong  
1782 topologies. 2016.
- 1783 [86] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American  
1784 Mathematical Soc., 2017.
- 1785 [87] Shige Peng. Nonlinear expectation theory and stochastic calculus under knightian uncertainty.  
1786 In *Real Options, Ambiguity, Risk and Insurance*, pages 144–184. IOS Press, 2013.
- 1787 [88] Nikhil R Pal, James C Bezdek, and Rohan Hemasinha. Uncertainty measures for evidential  
1788 reasoning i: A review. *International Journal of Approximate Reasoning*, 7(3-4):165–183,  
1789 1992.
- 1790 [89] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for  
1791 computer vision? *Advances in neural information processing systems*, 30, 2017.
- 1792 [90] Fabio Cuzzolin. Uncertainty measures: A critical survey. *Information Fusion*, page 102609,  
1793 August 2024. doi: 10.1016/j.inffus.2024.102609. URL [https://linkinghub.elsevier.  
1794 com/retrieve/pii/S1566253524003877](https://linkinghub.elsevier.com/retrieve/pii/S1566253524003877).
- 1795 [91] Joaquín Abellán and George J Klir. Additivity of uncertainty measures on credal sets. *International  
1796 Journal of General Systems*, 34(6):691–713, 2005.
- 1797 [92] Radim Jiroušek and Prakash P Shenoy. A new definition of entropy of belief functions in the  
1798 dempster–shafer theory. *International Journal of Approximate Reasoning*, 92:49–65, 2018.
- 1799 [93] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of  
1800 Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison.  
1801 In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*,  
1802 pages 548–557. PMLR, August 2022. URL [https://proceedings.mlr.press/v180/  
1803 hullermeier22a.html](https://proceedings.mlr.press/v180/hullermeier22a.html). ISSN: 2640-3498.
- 1804 [94] Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the Volume of a Credal Set a Good Mea-  
1805 sure for Epistemic Uncertainty?, June 2023. URL <http://arxiv.org/abs/2306.09586>.  
1806 arXiv:2306.09586 [cs, stat].
- 1807 [95] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for  
1808 imprecise probabilities. *Risk, Decision and Policy*, 5(2):165–181, 2000.

- 1809 [96] Fabio G Cozman. Sets of probability distributions and independence. *SIPTA Summer School*  
1810 *Tutorials, July 2-8, Montpellier, France*, 2008.
- 1811 [97] Joaquín Abellán, George J Klir, and Serafín Moral. Disaggregated total uncertainty measure  
1812 for credal sets. *International Journal of General Systems*, 35(1):29–44, 2006.
- 1813 [98] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of  
1814 credal uncertainty in machine learning: A critical analysis and empirical comparison. In  
1815 *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.
- 1816 [99] Peter J. Veazie. When to Combine Hypotheses and Adjust for Multiple Tests. *Health*  
1817 *Services Research*, 41(3p1):804–818, June 2006. ISSN 0017-9124, 1475-6773. doi: 10.1111/  
1818 j.1475-6773.2006.00512.x. URL [https://onlinelibrary.wiley.com/doi/10.1111/j.](https://onlinelibrary.wiley.com/doi/10.1111/j.1475-6773.2006.00512.x)  
1819 [1475-6773.2006.00512.x](https://onlinelibrary.wiley.com/doi/10.1111/j.1475-6773.2006.00512.x).
- 1820 [100] Sébastien Destercke. Handling bipolar knowledge with imprecise probabilities. *International*  
1821 *Journal of Intelligent Systems*, 26(5):426–443, March 2012. doi: 10.1002/int.20475. Publisher:  
1822 Wiley.
- 1823 [101] Andrey G. Bronevich and Natalia S. Spiridenkova. Some Characteristics of Credal Sets and  
1824 Their Application to Analysis of Polls Results. *Procedia Computer Science*, 122:572–578,  
1825 2017. ISSN 18770509. doi: 10.1016/j.procs.2017.11.408.
- 1826 [102] Michele Caprio and Ruobin Gong. Dynamic precise and imprecise probability kinematics. In  
1827 *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and*  
1828 *Applications*, pages 72–83. PMLR, July 2023. URL [https://proceedings.mlr.press/](https://proceedings.mlr.press/v215/caprio23a.html)  
1829 [v215/caprio23a.html](https://proceedings.mlr.press/v215/caprio23a.html). ISSN: 2640-3498.
- 1830 [103] Michele Caprio and Sayan Mukherjee. Ergodic theorems for dynamic imprecise probability  
1831 kinematics. *International Journal of Approximate Reasoning*, 152:325–343, 2023.
- 1832 [104] Inés Couso, Laura Garrido, and Luciano Sánchez. Similarity and dissimilarity measures  
1833 between fuzzy sets: a formal relational study. *Information Sciences*, 229:122–141, 2013.
- 1834 [105] Susana Montes, Susana Díaz, and Davide Martinetti. Divergence measures: from uncertainty  
1835 to imprecision. *The Mathematics of the Uncertain: A Tribute to Pedro Gil*, pages 675–684,  
1836 2018.
- 1837 [106] Robert Hable. A minimum distance estimator in an imprecise probability model—computational  
1838 aspects and applications. *International journal of approximate reasoning*, 51(9):1114–1128,  
1839 2010.
- 1840 [107] Walter L Perry and Harry E Stephanou. Belief function divergence as a classifier. In *Proceed-*  
1841 *ings of the 1991 IEEE International Symposium on Intelligent Control*, pages 280–285. IEEE,  
1842 1991.
- 1843 [108] Samuel S Blackman and Robert Popoli. Design and analysis of modern tracking systems. (*No*  
1844 *Title*), 1999.
- 1845 [109] Anne-Laure Jousselme, Dominic Grenier, and Éloi Bossé. A new distance between two bodies  
1846 of evidence. *Information fusion*, 2(2):91–101, 2001.
- 1847 [110] George Klir and Mark Wierman. *Uncertainty-based information: elements of generalized*  
1848 *information theory*, volume 15. Springer Science & Business Media, 1999.
- 1849 [111] Fuyuan Xiao, Weiping Ding, and Witold Pedrycz. A generalized  $f$ -divergence with appli-  
1850 cations in pattern classification. *IEEE Transactions on Knowledge and Data Engineering*,  
1851 2025.
- 1852 [112] Marta Catalano and Hugo Lavenant. Hierarchical integral probability metrics: A distance on  
1853 random probability measures with low sample complexity. *arXiv preprint arXiv:2402.00423*,  
1854 2024.
- 1855 [113] Joaquin Abellan and Serafin Moral. Maximum of entropy for credal sets. *International journal*  
1856 *of uncertainty, fuzziness and knowledge-based systems*, 11(05):587–597, 2003.

- 1857 [114] Yusuf Sale, Michele Caprio, and Eyke Höllermeier. Is the volume of a credal set a good  
1858 measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*, pages 1795–1804.  
1859 PMLR, 2023.
- 1860 [115] Yarin Gal et al. Uncertainty in deep learning. 2016.
- 1861 [116] Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying  
1862 aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual  
1863 information appropriate measures? In Robin J. Evans and Ilya Shpitser, editors, *Proceedings*  
1864 *of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Pro-*  
1865 *ceedings of Machine Learning Research*, pages 2282–2292. PMLR, 31 Jul–04 Aug 2023. URL  
1866 <https://proceedings.mlr.press/v216/wimmer23a.html>.
- 1867 [117] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Fos-  
1868 ter, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. *arXiv preprint*  
1869 *arXiv:2412.20892*, 2024.
- 1870 [118] Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-order  
1871 uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023.
- 1872 [119] Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-Order Uncertainty  
1873 Quantification: A Distance-Based Approach, December 2023. URL [http://arxiv.org/](http://arxiv.org/abs/2312.00995)  
1874 [abs/2312.00995](http://arxiv.org/abs/2312.00995). arXiv:2312.00995 [cs, stat].
- 1875 [120] Arthur Hoarau, Sébastien Destercke, Yusuf Sale, Paul Hofman, and Eyke Hüllermeier. Mea-  
1876 sures of uncertainty: A quantitative analysis, 2025.
- 1877 [121] Iipm: Reproducibility code for neurips 2025 submission. [https://anonymous.4open.](https://anonymous.4open.science/r/IIPM_NeurIPS2025-B5E1)  
1878 [science/r/IIPM\\_NeurIPS2025-B5E1](https://anonymous.4open.science/r/IIPM_NeurIPS2025-B5E1), 2025.
- 1879 [122] Siu Lun Chau, Jean-Francois Ton, Javier González, Yee Teh, and Dino Sejdinovic. Bayesimp:  
1880 Uncertainty quantification for causal data fusion. *Advances in Neural Information Processing*  
1881 *Systems*, 34:3466–3477, 2021.
- 1882 [123] Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional downscaling with  
1883 gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825,  
1884 2021.
- 1885 [124] Joaquín Abellán and Manuel Gómez. Measures of divergence on credal sets. *Fuzzy Sets and*  
1886 *Systems*, 157(11):1514–1531, June 2006. ISSN 01650114. doi: 10.1016/j.fss.2005.11.021.
- 1887 [125] Estimation of obesity levels based on eating habits and physical condition [dataset], 2019.  
1888 URL <https://doi.org/10.24432/C5H31Z>. UCI Machine Learning Repository.
- 1889 [126] E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits [dataset], 1998. URL  
1890 <https://doi.org/10.24432/C50P49>.
- 1891 [127] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
1892 2009.
- 1893 [128] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan:  
1894 Towards deeper understanding of moment matching network. *Advances in neural information*  
1895 *processing systems*, 30, 2017.
- 1896 [129] Christopher Bülte, Yusuf Sale, Timo Löhr, Paul Hofman, Gitta Kutyniok, and Eyke Hüller-  
1897 meier. An axiomatic assessment of entropy-and variance-based uncertainty quantification in  
1898 regression. *arXiv preprint arXiv:2504.18433*, 2025.
- 1899 [130] Sébastien Destercke, Didier Dubois, and Eric Chojnacki. Unifying practical uncertainty  
1900 representations–i: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49  
1901 (3):649–663, 2008.

- [131] F. J. Giron and S. Rios. Quasi-Bayesian Behaviour: A more realistic approach to decision making? *Trabajos de Estadística Y de Investigación Operativa*, 31(1):17–38, February 1980. ISSN 0041-0241. doi: 10.1007/BF02888345. URL <http://link.springer.com/10.1007/BF02888345>.
- [132] Leonard J. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.
- [133] Jon Williamson. *In Defence of Objective Bayesianism*. Oxford University Press, April 2010. ISBN 978-0-19-922800-3. doi: 10.1093/acprof:oso/9780199228003.001.0001.
- [134] Hamed Rahimian and Sanjay Mehrotra. Distributionally Robust Optimization: A Review. *Open Journal of Mathematical Optimization*, 3:1–85, July 2022. ISSN 2777-5860. doi: 10.5802/ojmo.15. URL <http://arxiv.org/abs/1908.05659>. arXiv:1908.05659 [cs, math, stat].
- [135] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple Source Adaptation and the Renyi Divergence, May 2012. arXiv:1205.2628 [cs, stat].
- [136] Simon Föll, Alina Dubatovka, Eugen Ernst, Siu Lun Chau, Martin Maritsch, Patrik Okanovic, Gudrun Thaeter, Joachim Buhmann, Felix Wortmann, and Krikamol Muandet. Gated domain units for multi-source domain generalization. *Transactions on Machine Learning Research*, 2023.
- [137] Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *International Conference on Learning Representations*, 2025.
- [138] Fabio Cuzzolin. Uncertainty measures: The big picture, April 2021. URL <http://arxiv.org/abs/2104.06839>. arXiv:2104.06839 [cs, math, stat].
- [139] Ibrahim Ghafir, Konstantinos G Kyriakopoulos, Francisco J Aparicio-Navarro, Sangarapillai Lambotharan, Basil AsSadhan, and Hamad BinSalleeh. A basic probability assignment methodology for unsupervised wireless intrusion detection. *IEEE Access*, 6:40008–40023, 2018.
- [140] Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [141] Muhammad Mubashar, Shireen Kudukkil Manchingal, and Fabio Cuzzolin. Random-set large language models. *arXiv preprint arXiv:2504.18085*, 2025.
- [142] Freddy Delbaen. Convex games and extreme points. *Journal of Mathematical Analysis and Applications*, 45(1):210–233, 1974.
- [143] Ignacio Montes, Enrique Miranda, and Paolo Vicig. 2-monotone outer approximations of coherent lower probabilities. *International Journal of Approximate Reasoning*, 101:181–205, 2018.
- [144] Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*, volume 130. Springer, 2021.
- [145] Stefan Aeberhard and Michele Forina. Wine [dataset]. <https://doi.org/10.24432/C5PC7J>, 1992. UCI Machine Learning Repository.
- [146] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.
- [147] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- [148] Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019.

- 1950 [149] B-E. Chérif-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maxi-  
1951 mum mean discrepancy. In *Proceedings of The 2nd Symposium on Advances in Approximate*  
1952 *Bayesian Inference (AABI)*, pages 1–21, 2020.
- 1953 [150] Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat.  
1954 Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71,  
1955 2021.
- 1956 [151] D. Sejdinovic. An overview of causal inference using kernel embeddings. *arXiv:2410.22754*,  
1957 2024.
- 1958 [152] Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. Rkhs-shap: Shapley values  
1959 for kernel methods. *Advances in neural information processing systems*, 35:13050–13063,  
1960 2022.
- 1961 [153] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learn-  
1962 ing from distributions via support measure machines. In *Advances in Neural Information*  
1963 *Processing Systems 25*, pages 10–18. 2012.
- 1964 [154] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning  
1965 theory for distribution regression. *Journal of Machine Learning Research*, 17(1):5272–5311,  
1966 2016.
- 1967 [155] Antonin Schrab and Ilmun Kim. Robust kernel hypothesis testing under data corruption. *arXiv*  
1968 *preprint arXiv:2405.19912*, 2024.
- 1969